

Statistical Learning on a Complex Metabolomic Dataset

Xiaodong Lin,^{*} Susan J. Simmons[†],
Chris Beecher[‡], Young Truong[§], and S. Stanley Young[¶]

Abstract

The post-genomic era has been driven by the development of technologies that allow the quantitative exploration of whole organisms at the molecular level. Metabolomics is concerned with the measurement of large sets of metabolites over biological samples using analytical techniques such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). The profiling of metabolites may provide a more comprehensive view of cellular control mechanisms in man and animals, and raise the possibility of identifying surrogate markers of disease. Advanced statistical tools are then employed to recover and interpret the information contained in the large data sets that are generated. Herein, we explore a complex metabolomic data set using a variety of statistical learning tools including recursive partitioning (RP), support vector machine (SVM) and random forest (RF). Comparing the results from random forest, feature selection SVM and recursive partitioning, we select an important set of metabolites for disease classification. The metabolites identified are linked to the known biochemistry function of the particular disease under investigation. Furthermore, we compare the outliers chosen by the RF to those identified by RP.

^{*}University of Cincinnati, Cincinnati, OH, 45221 linxd@math.uc.edu

[†]University of North Carolina, Wilmington, NC 28403 simmonssj@uncw.edu

[‡]Metabolon, Research Triangle Park, NC 27709 cbeecher@metabolon.com

[§]University of North Carolina, Chapel Hill, NC 27599 truong@bios.unc.edu

[¶]National Institute of Statistical Sciences, Research Triangle Park, NC 27709
young@niss.org

1 introduction

With the advent of the post-genomic era, biological and medical research has witnessed an explosion in strategies that provide an integrative view of the molecular regulation of cells and whole organisms. These advances have been driven by the development of novel technologies that can analyze global sets of biological products. In the emerging field of metabolomics, changes in populations of low-molecular-weight metabolites can be measured under a given set of conditions (Fiehn et al. 2000). In man and animals, metabolic profiles can be regarded as important indicators of normal phenotype and pathology, and offer the possibility of identifying surrogate biomarkers of disease states. Body fluids such as plasma and urine contain hundreds of small molecules and there may be as many as 20,000 metabolites in the plant kingdom (Fiehn, 2002) which introduces a considerable analytical challenge. Unlike genomics and, in theory, proteomics, where essentially just one class of compound is analysed, metabolomic strategies have to detect a broad spectrum of molecules with diverse properties (Weckwerth, 2003).

Even though the dynamic nature of metabolites makes them difficult to measure, recent advances in technology allows robust quantification of the concentrations of hundreds of metabolites from a biological sample. These new techniques offer insight into the dynamic interactions in metabolomic pathways. Furthermore, patterns of metabolites can be used to identify biomarkers of specific disease. Thus, how to analyze these complex data set poses an imminent challenge for researchers. In a typical metabolomic data set, the number of metabolites measured is usually much larger than the number of biological samples. Many of the metabolites are highly correlated, and some of them may be regarded as noise variables. Thus, it is of great challenge to identify those metabolites that are essential for disease classification.

Herein, we present analysis results using many state of the art statistical learning tools to select the important metabolites for an interesting metabolomic dataset. Our goal is to use these tools in an exploratory manner to identify metabolites that classify samples into consistent groups that

correspond to their known biological classification by experts. In particular, we apply three classification tools, namely, random forest (RF) (Breiman 2001), recursive partitioning (RP) and support vector machine(SVM) on this data set. Some preliminary analysis results on this dataset have been reported in Simmons et al. (2004) and Truong et al. (2004). In this paper, we compare the classification results of those methods which perform classification and variable selection simultaneously (RF, L-1 SVM (Mangasarian et al. 2002) and SCAD SVM (Zhang et al. 2005)) with those do not (standard SVM). We have found that in general, the former give superior performances. In addition to the selection of important metabolites, outlier detection is also very important in analyzing metabolomic data sets. Random forest has built in procedures for detecting outliers. Most of the outliers we obtain by applying RF match with those we found through RP.

The paper is organized in the following manner. Section 2 describes the data set in greater detail. Subsequently, in Sections 3, 4, and 5, we discuss RF, SVM and RP and the corresponding analyses. The comparisons and findings are reported in Section 6. Concluding remarks and some interesting research problems are described in Section 7.

2 The data

Our data consists of measurements of 317 metabolites for the blood sample of 63 subjects. There are two primary groups, 31 diseased(ALS disease) and 32 healthy individuals. Among the disease group, 22 subjects are diagnosed with a specific disease state and not taking medication, and 9 diseased subjects taking medication. The samples were analyzed using an ESA 16-channel Coularray detector. The Coularrays have very low detection limits and are able to detect metabolites with low molecular weight. This detector has a technical robustness that is comparable, possibly even better than conventional mRNA. The data contains blanks where the metabolic concentrations were below detection limits. The data set was collected to test the analytical equipment and to ascertain if metabolic information can provide insight to the diseased status for individuals with ALS. A sample of blood

will be taken from a control and a target population. The separation and mass spectrometry process should give a profile of metabolic products that is distinct for the target population. If this can be done, then there is an opportunity for better diagnosis.

This is a typical high dimensional and low sample size data set. Recently, many statistical learning methods have been proposed for analyzing this kind of data. Promising analysis results depend on the careful selection of important features and thorough consideration of outliers. In the following Sections, we will describe several popular methods including the RP, RF, standard SVM and two recent approaches: L-1 SVM and SCAD-SVM. We will compare the results and report the findings.

3 Random forest

The random forest algorithm (RF), created by Leo Breiman (2001), constructs a forest of tree classifiers which are used to predict the classification of a test set or training set. Each tree in the forest is constructed using a bootstrap sample of approximately two-thirds of the observations in the training set. At each node, a number of variables are randomly selected and used to create the next split. This procedure is repeated for each tree in the forest. Observations are classified by running the observation down each tree and allowing each tree to vote for which class the observation belongs. The observation is classified as the class which receives the most votes from the tree classifiers.

The RF algorithm provides an internal, unbiased estimate of the classification error rate (Breiman, 2001). Observations not included in the bootstrap sample to create a tree classifier are referred to as out-of-bag samples for that tree. Each out-of-bag observation is dropped down the tree classifier and its classification is recorded. The percent of observations misclassified provides an estimate for the out-of-bag error rate.

3.1 Classification error

Using the primary disease status information of diseased versus not diseased as the classification, we create a random forest via the FORTRAN code developed by Leo Breiman and Adele Culter. The mtry variable used in construction of the forest was 8. The error rate for a random forest created using all of the variables was identical to the error rate of a random forest created using only the most significant 20 variables. Therefore, we focus our discussions and results on the 20 variable model.

The out-of-bag error rate is 7.94% with the confusion matrix illustrated in Table 1.

Table 1: Classification error for two classes case.

	Class one	Class two
Class one	29	2
Class two	3	29

3.2 Variable selection

Feature selection or variable selection in RF is determined by using the out-of-bag observations for the tree classifiers. The out-of-bag observations are run down the tree and the number of votes for the correct classes is recorded. One variable in the out-of-bag observations is permuted and the observations are again run down the tree classifier with the number of votes for the correct classes recorded. The number of votes for the correct classes for the permuted data is subtracted from the number of votes for the correct classes from the untouched data. The average of this number across all trees is the raw importance score. This is repeated for each variable. A standardized importance score is calculated by dividing the raw importance score by its standard error. The 20 most important features along with the raw importance score and the standardized importance score is illustrated

in Table 2.

Table 2: Top 20 metabolites selected by important score.

Metabolite	Raw importance score	Standardized importance score
3-91.1	6.294	12.798
12-78.0	4.937	10.795
14-80.3	4.047	10.329
5-80.2	3.679	9.890
12-12.2	2.182	8.099
5-89.2	1.778	7.640
9-7.1	1.787	7.440
11-54.1	1.481	7.100
16-89.8	1.319	6.953
8-79.7	1.580	6.434
14-20.1	1.079	6.013
16-34.9	1.030	5.542
4-91.9	.933	5.098
4-37.9	.625	4.507
6-66.1	.677	4.265
16-97.5	.619	4.234
3-78.2	.472	3.897
12-92.2	.347	3.746
12-92.2	.500	3.343
3-82.7	.324	2.783

3.3 Outlier detection

Outliers are detected in RF by calculating a proximity measure among all observations. The proximity matrix is an N by N matrix displaying the proximities of observation i with observation j for $i, j = 1, 2, \dots, n$. The diagonal values of this matrix are the number of trees in the forest and the off-diagonal elements are calculated by the proximity of observation i with observation j . If two observations are in the same terminal node of a tree, the proximity is increased by one. Therefore, if two observations are in the same terminal node for all trees in the forest, the proximity between these two observations would be equal to the number of trees in the forest.

Outliers are observations with low proximities to other observations in the same class. The outlier measure for each of the 63 observations is illustrated in Figure 1. As shown, the four outliers are X16, X29, X57 and X58.

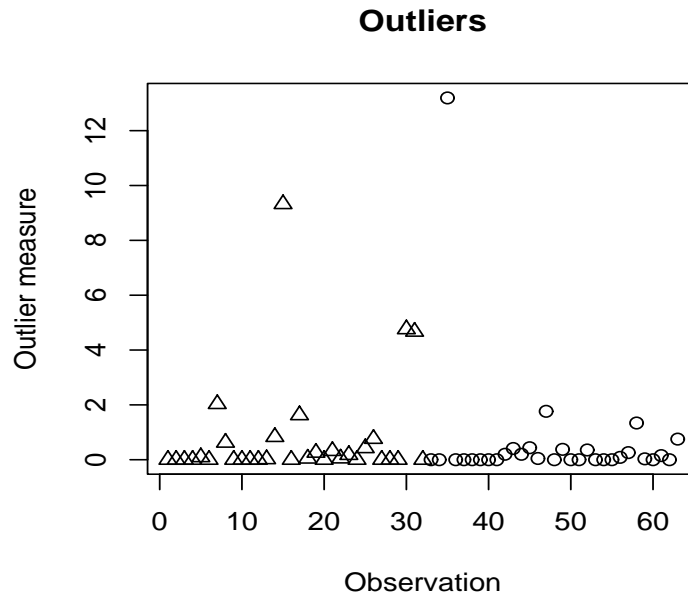


Figure 1: Outlier measure determined by RF.

4 Support Vector Machine

4.1 Standard SVM

Support vector machine (SVM) (Vapnik, 1995; Cristianini and Shawe-Taylor, 1999) is a large-margin classifier which separates two classes by maximizing the margin between them. It has demonstrated very good performances in classifying high dimensional and low sample size data. Given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x} = (x_1, \dots, x_d)$ is the d dimensional covariate vector

and y denotes the class label, the standard SVM find $f(\mathbf{x})$ to maximize

$$\frac{1}{n} \sum_{i=1}^n [1 - y_i f(\mathbf{x}_i)]_+ + \lambda \|\mathbf{w}\|^2, \quad (1)$$

where b is a constant and \mathbf{w} is the directional vector. In general, the SVM solution depends only on a small subset of the training samples called “support vectors”. When the function f is linear in \mathbf{x} , we obtain a linear SVM. When we choose the f to be polynomial functions or Gaussian kernels, we obtain corresponding non-linear SVM.

4.2 Feature selection SVM

The solution of the standard SVM utilizes all the input covariates without discrimination. A typical high dimensional low sample size dataset may contain many noisy and redundant variables. Without selecting the important variables, the standard SVM may give misleading classification results. Hastie et al. (2001) gave an illustrative example on this. Several approaches for variable selection in SVM have been proposed. In this paper, we will study two of these approaches, and their performances on the metabolomic dataset.

In the statistics literature, the linear regression model with the L_1 penalty is known as the LASSO. Tibshirani (1996) gave a thorough study of the method for variable selection. Bradley and Mangasarian (1998) proposed the L_1 SVM which solves

$$\min_{b, \mathbf{w}} \frac{1}{n} \sum_{i=1}^n [1 - y_i(b + \mathbf{w} \cdot \mathbf{x}_i)]_+ + \lambda \sum_{j=1}^d |w_j|. \quad (2)$$

The L_1 penalty in (2) gives a soft-thresholding rule and yields a directional vector $\hat{\mathbf{w}}$ with many zero components, thus achieves variable selection for SVM.

Fan and Li (2001) showed that, in the linear regression setting, the L_1 penalty produces biased solutions for large coefficients. Similar situation arises in the SVM context. Zhang et al. (2005) proposed the SCAD-SVM

which utilizes a non-convex penalty function in replacing the L_1 penalty.

The SCAD penalty function is defined as

$$p_\lambda(w) = \begin{cases} \lambda|w| & \text{if } |w| \leq \lambda, \\ -\frac{(|w|^2 - 2a\lambda|w| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |w| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |w| > a\lambda, \end{cases} \quad (3)$$

where $a > 2$ and $\lambda > 0$ are two tuning parameters. The function p_λ is symmetric and a quadratic spline function with two knots at λ and $a\lambda$. In Figure 2, we plot the SCAD function with $a = 3$ and $\lambda = 0.4$. Except its singularity at the origin, the function $p_\lambda(w)$ has a continuous first-order derivative.

The SCAD function has the same form as the L_1 penalty for small coefficients. However, for large coefficients, the SCAD always applies a constant penalty while the L_1 penalty increases linearly as the coefficient increases. It is this distinct feature that guards the SCAD penalty against causing possible biases for estimating large coefficients.

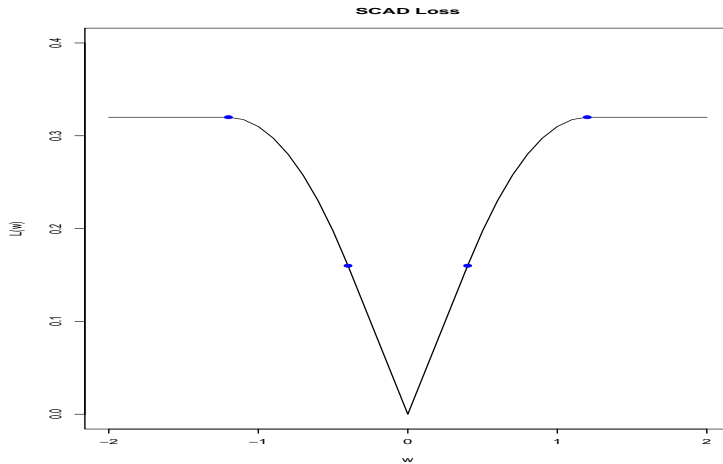


Figure 2: SCAD loss functions with $\lambda = 0.4$, $a = 3$.

4.3 Data analysis

In our analysis, we compare the linear SVM, the nonlinear SVM using polynomial and Gaussian Kernels, the L_1 SVM and the SCAD SVM. For the first three cases, we use the OSU SVM package (www.ece.osu.edu/maj/osu-svm) and the code of Fung and Mangasarian (2004) for the L_1 SVM. The classification errors reported are based on the leave one out cross validation. For SCAD-SVM, in each training set with 62 samples, cross validation is performed within the training set to tune the parameter λ .

Table 3 shows the average leave-out-one cross validation error and the number of metabolites selected by each method. Among the four methods, the SCAD SVM gives the smallest cross validation error 0.143 while the standard SVM performs worst. Moreover, the SCAD SVM selects 18 important metabolites out of 317 and the L_1 SVM selects 32 metabolites. Therefore the SCAD SVM gives the highest classification accuracy while using the fewest metabolites. This method may have great implications on metabolic studies, since one main issue from biological aspects is to identify which metabolites are more relevant to the occurrence of a certain disease. Among the 32 metabolites selected by L_1 SVM and the 18 metabolites selected SCAD-SVM, the set 5 – 80.2, 12 – 12.2, 14 – 80.3, 5 – 65.2 and 4 – 91.9 are commonly chosen.

Table 3: Cross validation error and the number of metabolites selected for metabolism data.

	Test error	Number of selected metabolites
Linear SVM	0.190 (0.013)	
Poly SVM	0.238 (0.012)	
Gaussian SVM	0.286 (0.017)	
L_1 SVM	0.174 (0.012)	32
SCAD SVM	0.143 (0.020)	18

5 Recursive Partitioning

Recursive partitioning is a tree-based classification algorithm that searches through all of the variables and identifies the variables producing the best or most significant split at each node. A significant split is determined by the p-value from either an F-test or t-test. Two-way splits are determined via a t-test, while multi-way splits are decided by an F-test. The FIRMPLUSTM algorithm, developed by Douglas Hawkins, will search for the optimal split for each variable. The criterion used to determine a significant split is based on the Bonferroni p-value, adjusting for the number of variables, and an adjusted p-value, accounting for the segmentation of each variable. One feature of RP is its ability to retain missing values throughout the analysis. Missing values are treated as possible values a variable can assume and therefore are kept as such. This feature is desirable when missing values are informative, as with metabolomic data.

5.1 Multiple tree analysis

The commercially available software, Partitionator was used to create multiple trees by the FIRMPLUSTM algorithm. An important difference in the construction of the multiple trees in RP to that of the RF is that only variables with significant splits at a node can be chosen. The random forest algorithm randomly selects variables from the entire variable list, where recursive partitioning randomly accepts variables regarded as significant at each node. The multiple trees provide information regarding interactions and correlations among the variables, or metabolites. In the construction of trees, two variables that are correlated should not appear in the same tree. Thus, once the variation due to one variable is removed (i.e. the node is split using this variable), the other variable should not exhibit a significant variation in this tree. The same idea can be extended to variables with an interaction. Two variables that interact should appear more often than by chance in the same tree.

Figure 3 illustrates this information. The numbers on the diagonal are the percent of all cases in the multiple trees that are split using that variable.

The upper triangular values are the percent of cases jointly split by the two variables. For example, notice that 14-80.3 and 16-86.1 jointly split 12% of all observations in the multiple trees. The lower diagonal values represent a standardized score of the upper triangular values.

RP does not perform any internal cross-validation analysis. Therefore, we manually reran the multiple tree analysis on several randomly selected subsets of observations. We randomly choose two-thirds of the observations and reconstructed the multiple trees. Unfortunately, we were unable to get an estimate of classification error from this method.

	5-80.2	16-86.1	14-80.3	8-79.7	5-65.2	9-7.1	11-54.1
5-80.2	0.18	0	0	0	0	0	0
16-86.1	-1.1	0.12	0.12	0	0	0	0
14-80.3	-1.4	5.4	0.17	0	0	0	0
8-79.7	-1.4	-1.1	-1.3	0.17	0	0	0
5-65.2	-1.4	-1.1	-1.3	-1.3	0.17	0	0
9-7.1	-1.4	-1.1	-1.3	-1.3	-1.3	0.16	0
11-54.1	-1.3	-1.1	-1.3	-1.3	-1.3	-1.3	0.15

Figure 3: Correlations and interactions from RP.

5.2 Important variable selection

For data sets with a relatively small number of observations, it is advantageous to build many trees using the RP algorithm. The construction of multiple trees provides information regarding important features, since those essential features should be used often. In our analysis, 1000 trees were constructed and six metabolites were identified as critical in the multiple trees. These metabolites were 5-80.2, 14-80.3, 8-79.7, 5-65.2, 9-7.1, 11-54.1, and 16-86.1. Given that the RP algorithm does not have an internal mechanism to validate the chosen features, we manually executed a cross-validation procedure by randomly selected a subset of approximately two-thirds of the observations and recreating the multiple trees. This was repeated several times with variables 5-80.2, 14-80.3 and 5.65.2 consistently present in all multiple trees.

5.3 Outlier detection

Although RP does not have a formal outlier detection algorithm, outliers in RP can be discovered by observing cases misclassified in the final nodes of the classification tree. In the metabolomic data set, recursive partitioning suggested observations X58, X16 and X43 are potential outliers.

6 Performance evaluation

A number of recent studies have utilized the statistical learning tools of RF, SVM, and RP. In this section, we compare the performance of these methods with respect to classification, variable selection, and outlier detection on a rich metabolomic dataset.

Classification

One of the fundamental goals of analyzing the metabolomic dataset is to correctly classify samples into the appropriate disease status based on their metabolite measurements. The dataset contains healthy and diseased subjects from which we would like to build generic learning rules to sepa-

rate them. Table 4 reports the classification error rates for RF and SVM. Unfortunately, the RP algorithm does not procure a classification error rate.

Table 4: Comparisons on two class classification rate.

Method	RF	SVM	RP
Classification error	0.0794	0.143	-

Apparently RF gives the best classification results among the three methods. Furthermore, RF can be applied to analyze multi-category problems. The disease group can be subdivided into two groups, one taking a prescribed medication for the disease and one which does not. In this situation, there are three classes and RF gives an OOB error rate of 12.7% as shown in Table 5.

Table 5: Three class classification by RF.

	Class one	Class two	Class three
Class one	32	1	4
Class two	0	21	3
Class three	0	0	2

Metabolite selection

It is known that for high dimensional classification problems, noise features and redundant variables can affect the classifier dramatically. Incorporating variable selection with classification can be beneficial. Analysis from two SVM methods reveals metabolites 5-80.2, 12-12.2, 14-80.3, 5-65.2 and 4-91.9 as important. All methods agree that metabolites 5-80.2 and 14-80.3 are important features in distinguishing the disease status of the subjects. Furthermore, metabolites 12-12.2 and 4-91.9 are designated to be important features by both of the SVM methods and RF. RP and the SVM methods also agree that metabolite 5-65.2 is an important feature. As illustrated, some metabolites are common to all three procedures, while others only agree on a subset of methods.

Outlier detection

Four samples, namely X16, X29, X57 and X58 are detected as outliers by RF. Interestingly, three of these four (X16, X29 and X58) are also misclassified in the confusion matrix in Table 1. Meanwhile, RP suggested observations X58, X16 and X43 are potential outliers. The two samples X16 and X58 are commonly selected. We are currently investigating the biological implications of these two samples being selected as outliers.

7 Conclusion

We have studied the classification performances of RF, SVM and RP on a rich metabolomic dataset. We have also examined these tools further in terms of selecting important metabolites for disease classification. The standard SVM methods perform similarly to those reported in (Truong et al., 2004), and the L_1 approach has a slightly improvement. The SCAD-SVM has a small reduction in the error rate while utilizing nearly half the number of metabolites used by the other SVM methods. We view this as a significant improvement. We have also observed that RF yields better classification over the SVM based methods using about the same number of metabolites. Moreover, RF provides a detailed analysis of the importance of these metabolites in terms of sensitivity as well as accuracy for classification. The former is very useful as it paves a pathway for further understanding of the relationship of these metabolites and the biochemistry of the disease.

References

- [1] Bradley, P. S. and Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. *In Proceeding of the 13th International Conference on Machine Learning*, 82-90, San Francisco, CA.
- [2] Breiman, L. (2001). Random forests. *Machine Learning*, 45 (1), 5-32.
- [3] Fan, J. and Li, R. Z. (2001). Variable selection via penalized likelihood. *Journal of the American Statistical Association*, 96, 1348-1360.
- [4] Fiehn, O. (2002). Metabolomics the link between genotype and phenotype. *Plant Molecular Biology*, 48, 155-171.
- [5] Fiehn, O., Kopka, J., Drmann, P., Altmann, T., Trethewey, R.N., Willmitzer, L. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnology* 18, 1157-1161.
- [6] Fung, G. and Mangasarian, O. L. (2004). A feature selection newton method for support vector machine classification. *Computational Optimization and Applications Journal*, 28(2), 185-202.
- [7] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: data mining, inference, and prediction*. Springer, New York.
- [8] Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, B*, 58, 267-288.
- [9] Simmons, S. J., Lin, X., Beecher, C., Truong, Y. and Young, S. S. (2004). Active and Passive Learning to Explore a Complex Metabolism Data Set. *In Proceedings of 2004 Meeting of International Federation of Classification Societies*.
- [10] Truong, Y., Lin, X., Beecher, C., Cutler, A., and Young, S. S.(2004). Learning Metabolomic Datasets with Random Forests and Support Vec-

tor Machines. *In Proceedings of Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

- [11] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.
- [12] Weckwerth, W. (2003) Metabolomics in Systems Biology. *Annual Review of Plant Biology*, 54, 669-689.
- [13] Zhang, H., Ahn, J., Lin, X., and Park, C. (2005). Variable Selection for SVM using Nonconcave Penalty. *In preparation.*
- [14] FIRMPPlusTM: Golden Helix, Inc., Bozeman, MT (www.goldenhelix.com).